مثال هایی از الگوریتم EM

برنارد فلوری ، آلیس زوپ[۱] ترجمهٔ عباس پرچمی[۲]

چکیده

الگوریتم EM ابزاری مفید در حل بسیاری از مسائل با داده‌های ناقص می‌باشد و در غالب کتاب های درسی آماری ارائه می‌شود. وقتی داده‌های بقا از توزیع نمایی پیروی کنند و بعضی از داده‌ها از سمت راست سانسور شده باشند، الگوریتم EM راه حل مناسبی برای برآورد حداکثر درستنمایی[۳] می‌باشد. امّا اگر این داده‌های طول عمر از سمت چپ هم سانسور شده باشند و یا اگر از توزیع یکنواخت پیروی کنند، آیا باز هم الگوریتم EM کارایی دارد؟ در این مقاله دو مثال ارائه شده است که تعمیمی از یک مثال معروف دربارهٔ الگوریتم EM می‌باشد. این مثالها، به دو نکته کارایی الگوریتم و خاصیت خودسازگاری آن تاکید می‌کنند.

واژه‌های کلیدی: الگوریتم EM ، توزیع نمایی، داده‌های ناقص، برآورد بیشترین درستنمایی، توزیع یکنواخت.

۱. مقدمه

الگوریتم EM که توسط دمپستر، لیرد و رابین[۴] در سال ۱۹۷۷ و توسط مک لاچلان و کریشنان[۵] در سال ۱۹۹۷ ارائه شد، ابزاری قدرتمند برای برآورد حداکثر درستنمایی با داده‌های ناقص می‌باشد. در اینجا معنی کلمهٔ « ناقص» حالتی کلی دارد و در موقعیتهای متفاوت می‌تواند به معانی گوناگونی مانند: داده‌های گم شده، مؤلفه‌های مجهول، مشاهدات سانسور شده، متغیرهای نهفته و نظایر آنها اشاره داشته باشد. شرح مختصری از الگوریتم EM در زیر ارائه می‌شود.

فرض کنید $\mathbf{Y}$ بردار داده‌های مشاهده شده و $\mathbf{X}$ بردار داده‌های نامعلوم باشند. همچنین فرض کنید θ پارامتر مجهول است که می‌خواهیم برآورد شود و $l_c(\theta;\mathbf{Y},\mathbf{X})$ لگاریتم تابع درستنمایی بر اساس کلیهٔ داده‌ها می‌باشد، که به ازای تمام مقادیر ممکن θ در فضای پارامتر Ω ، تعریف می‌شود.

الگوریتم با یک مقدار اولیهٔ $\theta^{(0)} \in \Omega$ آغاز می‌شود و دو مرحلهٔ زیر را تا رسیدن به همگرایی، تکرار می‌کند:

مرحلهٔ E[۶]:

محاسبهٔ $l^{(j)}(\theta) = E_{\mathbf{X}|\mathbf{Y},\theta^{(j-1)}}[l_c(\theta;\mathbf{Y},\mathbf{X})]$ که امید ریاضی با توجه به داده‌های گم شدهٔ $\mathbf{X}$ به شرط داده‌های مشاهده شدهٔ $\mathbf{Y}$ گرفته می‌شود و باید توجه شود که مقدار $\theta^{(j-1)}$ در این امید ریاضی جایگذاری می‌شود.

مرحلهٔ M[۷]: یافتن $\theta^{(j)} \in \Omega$ بقسمی که $l^{(j)}(\theta)$ بیشینه شود.

تکرار دو مرحلهٔ فوق به ازای $j = ۱,۲,...$ ، منجر به همگرایی دنبالهٔ $\theta^{(۱)}, \theta^{(۲)}, ...$ در ماکزیمم موضعی لگاریتم درستنمایی کلیّهٔ داده‌ها می‌شود. لازم به ذکر است که تحت شرایط خاصی این دنباله به همگرایی نمی‌رسد، برای جزئیات بیشتر به [۴] رجوع شود.

از جمله کاربردهای رایج این الگوریتم، حل مسائلی با داده‌های گم شده، یافتن نمای توزیع پسین در چهارچوب بیز[۳]، تشخیص مدلهای آمیخته و کاربردهایی در داده‌های گروه‌بندی شده، سانسور شده و یا بریده شده، می‌باشند؛ برای توضیح بیشتر به [۱]و [۲] مراجعه شود.

---

۱ - Bernard Flury and Alice Zoppe

۲ - دانشجوی کارشناسی ارشد آمار دانشگاه شهید باهنر کرمان

۳ - Maximum Likelihood Estimation

۴ - Dempster, Laird and Rubin

۵ - McLachlan and Krishnan

۶ - Expectation

۷ - Maximization

مثال ۱) فرض کنید طول عمر لامپ های مربوط به یک کارخانه دارای توزیع نمایی با میانگین مجهول θ باشد. از این کارخانه تعداد M+N لامپ را به تصادف انتخاب کرده و در دو آزمایش مستقل شرکت داده می شود. آزمایشگر تعداد N لامپ را در آزمایش اوّل قرار داده و طول عمر تک‌تک آنها را به صورت $y_1,...,y_N$ ثبت می‌کند، در حالیکه در آزمایش دوم زمان $t > 0$ را در نظر گرفته و همه M لامپ را داخل آزمایش می‌کند و فقط لامپ هایی را که تا زمان t هنوز نسوخته‌اند مشخص می کند. بنابراین در آزمایش دوم طول عمر دقیق تک تک مشاهدات مشخص نیست و تنها مشخصه های $E_1, E_2, ..., E_M$ در دسترس هستند که در آن

$$E_i = \begin{cases} 1 & \text{اگر لامپ تا زمان } t \text{ سوخته نباشد} \\ 0 & \text{اگر لامپ قبل از زمان } t \text{ سوخته باشد} \end{cases}$$

$, \quad i = 1,..., \quad M$

حال سؤال اینجاست که با این داده ها $MLE(\theta)$ چقدر می‌شود؟

فرض کنید که $X_1, ..., X_M$ طول عمرهای (مشاهده نشده) لامپ های شرکت کننده در آزمایش دوم باشند و $Z = \sum_{i=1}^{M} E_i$ تعداد لامپ هایی باشد که در آزمایش دوم تا زمان t هنوز نسوخته‌اند. بنابراین ترکیب داده‌های مشاهده شده در دو آزمایش به صورت زیر است:

$$\mathbf{Y} = (Y_1,...,Y_N, E_1,...,E_M)$$

و داده‌های مشاهده نشده عبارتند از:

$$\mathbf{X} = (X_1,...,X_M)$$

همچنین تابع درستنمایی بر اساس کلیهٔ داده‌ها به صورت زیر است:

$$L(\theta;\mathbf{Y},\mathbf{X}) = \frac{1}{\theta^{M+N}} \exp\left(\frac{-(\sum_{i=1}^{N} Y_i + \sum_{i=1}^{M} X_i)}{\theta} \right)$$

بنابراین خواهیم داشت:

$$\begin{aligned} l_c(\theta;\mathbf{Y},\mathbf{X}) &= \ln L(\theta) \\ &= -(M+N)\log\theta - \frac{1}{\theta}(\sum_{i=1}^{N} Y_i + \sum_{i=1}^{M} X_i) \end{aligned} \quad (1)$$

حال $E(X_i|\mathbf{Y})$ را در دو حالت $\mathrm{E_i} = 0$ و $\mathrm{E_i} = 1$ محاسبه می‌کنیم. با توجه به این موضوع که پیشامد $\mathrm{E_i} = 1$ معادل $X_i > t$ می‌باشد، می‌توان نوشت:

$$f(x_i \mid x_i > t) = \frac{f(x_i)}{1-F(t)} = \frac{1}{\theta} e^{\frac{t}{\theta} - \frac{x_i}{\theta}}$$

$$\begin{aligned} E(X_i \mid E_i = 1) &= E(X_i \mid X_i > t) \\ &= \int_t^{\infty} x_i f(x_i \mid x_i > t) dx_i = t + \theta \end{aligned}$$

بطور مشابه در حالت $\mathrm{E_i} = 0$ خواهیم داشت:

$$\begin{aligned} E(X_i \mid E_i = 0) &= \int_0^t x_i f(x_i \mid x_i \le t) dx_i \\ &= \theta - \frac{t e^{t/\theta}}{1 - e^{t/\theta}} \end{aligned}$$

در نتیجه می‌توان نوشت:

$$E(X_i|\mathbf{Y}) = E(X_i|E_i) = \begin{cases} t+\theta & \mathrm{E_i} = 1 \\ \theta - \dfrac{t e^{t/\theta}}{1-e^{t/\theta}} & \mathrm{E_i} = 0 \end{cases} \quad (2)$$

حال می‌توان گفت که jامین مرحله از الگوریتم شامل جایگذاری $E(X_i|E_i)$ از رابطهٔ (۲)، بجای X_i در رابطهٔ (۱) می‌باشد. لذا خواهیم داشت:

$$\begin{aligned} l^{(j)}(\theta) = -(N+M)\log\theta - \frac{1}{\theta}[&N\bar{Y} + Z(t+\theta^{(j-1)}) \\ &+ (M-Z)(\theta^{(j-1)} - tp^{(j-1)})] \end{aligned} \quad (3)$$

که در آن $p^{(j)} = \dfrac{e^{-t/\theta^{(j)}}}{1-e^{-t/\theta^{(j)}}}$ و منظور از $\theta^{(j)}$، θی بدست آمده در jامین تکرار می‌باشد. در jامین تکرار مرحلهٔ M، $\theta^{(j)}$ را چنان برآورد می‌کنیم که رابطهٔ (۳) را بیشینه می‌سازد، به عبارت دیگر با فرض ثابت بودن $\theta^{(j+1)}$، برآورد بیشترین درستنمایی $\theta^{(j)}$ را بدست می‌آوریم:

$$\frac{\partial l^{(j)}(\theta)}{\partial\theta} = 0 \quad \Rightarrow \quad (4)$$

$$\begin{aligned} \theta^{(j)} &= f(\theta^{(j-1)}) \\ &\equiv \frac{N\bar{Y} + Z(t+\theta^{(j-1)}) + (M-Z)(\theta^{(j-1)} - tp^{(j-1)})}{N+M} \end{aligned}$$

حال می‌توان با انتخاب یک $\theta^{(0)} > 0$ معادلهٔ (۴) را تا رسیدن به همگرایی تکرار کرد. خاصیّت خودسازگاری الگوریتم وقتی آشکار می‌شود که $Z = M$ باشد (یعنی کلیهٔ لامپ هایی که در آزمایش دوّم شرکت داده‌ایم تا زمان t سالم باشند)، در این حالت به جواب معروف زیر می‌رسیم:

$$\hat{\theta} = \frac{N\bar{Y} + Mt}{N}$$

این مثال بدون استفاده از الگوریتم EM نیز قابل حل است. لگاریتم تابع درستنمایی توأم بر اساس داده‌های مشاهده شده به صورت زیر بدست می‌آید:

$$L(\theta) = \frac{1}{\theta^N} e^{-N\bar{Y}/\theta} \left(e^{-t/\theta}\right)^Z \left(1 - e^{-t/\theta}\right)^{M-Z}$$

$$l(\theta) = -N(\log\theta + \frac{\bar{Y}}{\theta}) - \frac{Zt}{\theta} + (M - Z)\log(1 - e^{-t/\theta}) \qquad (۵)$$

حال به کمک روشهای عددی استاندارد می‌توان θ را چنان برآورد کرد که لگاریتم تابع درستنمایی (۵) بیشینه گردد. در حالت خاص $Z = M$ ، معادلهٔ خودسازگار و نیز ماکزیمم (۵) بطور تحلیلی قابل حل است، و در این حالت روشهای ML[1] و EM جوابهای مشابهی خواهند داشت [۲] .

حال بهتر است عملاً ببینیم چگونه به کمک معادلهٔ (۴) و با انتخاب هر $\theta^{(0)}$ای به یک همگرایی می‌رسیم. مثال اوّل را با داده‌های زیر در نظر بگیرید، در آزمایش اوّل طول عمر N=۴ لامپ بترتیب ۲/۵ ، ۳ ، ۱ و ۷/۵ ثبت شده و در آزمایش دوّم که با M=۶ لامپ صورت گرفت و با فرض t=۶ مقادیر E_iها بترتیب ۱، ۰، ۰، ۰، ۱ و ۰ مشاهده شدند و لذا Z=۲ و $\bar{Y}$=۳/۵ بدست آمده و معادلهٔ (۴) به شکل زیر تبدیل خواهد شد:

$$\theta^{(j)} \equiv \frac{۲۶ + ۶\theta^{(j-1)} - \left(\frac{۲۴\exp(-۶/\theta^{(j-1)})}{1 - \exp(-۶/\theta^{(j-1)})}\right)}{۱۰}$$

جدول زیر تکرار الگوریتم و رسیدن به همگرایی آن را به ازای دو مقدار $\theta^{(0)}$ کاملاً متفاوت نشان می‌دهد:

j	$\theta_1^{(j)}$	$\theta_2^{(j)}$	J	$\theta_1^{(j)}$	$\theta_2^{(j)}$
۰	۵/۰۰۰۰	۶۰/۰۰۰	۶	۴/۴۲۰۸	۴/۴۳۰۲
۱	۴/۵۶۵۶	۱۵/۷۸۰	۷	۴/۴۲۰۷	۴/۴۲۳۱
۲	۴/۴۵۷۵	۶/۸۸۰۱	۸	۴/۴۲۰۷	۴/۴۲۱۳
۳	۴/۴۳۰۱	۵/۰۰۳۸	۹		۴/۴۲۰۸
۴	۴/۴۲۳۱	۴/۵۶۶۵	۱۰		۴/۴۲۰۷
۵	۴/۴۲۱۳	۴/۴۵۷۸	۱۱		۴/۴۲۰۷

که ملاحظه می‌شود به ازای هر دو مقدار $\theta_1^{(j)}$ و $\theta_2^{(j)}$ ، به یک همگرایی ($\hat{\theta}$=۴/۴۲۰۷) می‌رسیم. بدیهی است که انتخاب $\theta^{(0)}$=۵ انتخاب معقول تری است و ما را زودتر به همگرایی می‌رساند.

مثال ۲) در مثال اول فرض کنید طول عمر لامپ های کارخانه دارای توزیع یکنواخت در بازهٔ $[0,\theta)$ می‌باشند، بطوریکه $\theta > 0$ مجهول است. فرض کنید آزمایش گر مانند مثال اول، آزمایشات مشابهی انجام داده و دوباره در آزمایش دوم تنها مشخصه های $X_1,...,X_M$ را ثبت می‌کند.

می‌خواهیم از الگوریتم EM برای حل مسئلهٔ برآورد حداکثر درستنمایی استفاده کنیم. می‌دانیم $MLE(\theta)$ براساس داده‌های کامل (بطور فرضی کامل) برابر $\max\{X_{\max}, Y_{\max}\}$ می‌باشد، بطوریکه $Y_{\max}$ بیشترین طول عمر مشاهده شده و $X_{\max}$ بیشترین طول عمر مشاهده نشده باشد.

برای سادگی فرض کنید $Z \geq 1$ ، یعنی در پایان آزمایش دوم حداقل یک لامپ سالم مانده باشد بنابراین می‌توان مطمئن بود که $\theta \geq t$ است و می‌توان نوشت:

$$f(x_i \mid x_i > t) = \frac{\mathrm{I}_{[0,\theta]}(x_i)}{\theta - t}$$

$$E(X_i \mid E_i = 1) = E(X_i \mid X_i > t) = \int_t^\theta \frac{x_i}{\theta - t} dx_i = \frac{t + \theta}{2}$$

بطور مشابه داریم:

$$E(X_i \mid E_i = 0) = \frac{t}{2}$$

در نتیجه می‌توان نوشت:

$$E(X_i|E_i) = \begin{cases} \frac{t+\theta}{2} & E_i = 1 \\ \frac{t}{2} & E_i = 0 \end{cases}$$

$$l^{(j)}(\theta) = \theta^{-(N+M)} I_{[\max\{Y_{\max}, t/2, (t+\theta^{(j-1)})/2\}, \infty)}(\theta)$$
$$= \theta^{-(N+M)} I_{[\max\{Y_{\max}, (t+\theta^{(j-1)})/2\}, \infty)}(\theta)$$

و الگوریتم EM شامل تکرار سادهٔ معادلهٔ زیر خواهد بود:

$$\theta^{(j)} = f(\theta^{(j-1)}) \equiv \max\{Y_{\max}, \frac{1}{2}(t + \theta^{(j-1)})\} \qquad (۶)$$

با انتخاب یک $\theta^{(0)} > 0$ به ازای $j = 1,2,...$ معادلهٔ (۶) همگرا به $\hat{\theta} = \max\{Y_{\max}, t\}$ می‌شود، و این نتیجه با توجه به خاصیت خودسازگاری و حل معادلهٔ $\theta = f(\theta)$ برحسب $\hat{\theta}$ نیز بدست

[1] - Maximum Likelihood

می‌آید. مزیت اصلی این راه حل (الگوریتم EM) سادگی آن می‌باشد، ولی این راه حل اشتباه است، چرا؟ (بعد از ارائه راه حل صحیح، توضیح خواهیم داد).

تابع درستنمایی توأم مربوط به داده‌های بدست آمده از هر دو آزمایش به صورت زیر است:

$$L(\theta)=\theta^{-N}I_{[Y_{\max},\infty)}(\theta)\times\left(\frac{t}{\max(t,\theta)}\right)^{M-Z}\left(1-\frac{t}{\max(t,\theta)}\right)^{Z} \quad (۷)$$

ابتدا حالت خاص $Z=0$ را در نظر بگیرید که

$$L(\theta)=\theta^{-N}I_{[Y_{\max},\infty)}(\theta)\left(\frac{t}{\max(t,\theta)}\right)^{M}$$

تابعی نزولی از $\theta\geq Y_{\max}$ می‌باشد و بنابراین برآورد بیشترین درستنمایی برابر $\hat{\theta}=Y_{\max}$ می‌شود، سپس حالت $Z\geq 1$ را در نظر بگیرید که بر $\theta\geq t$ دلالت دارد. به ازای $\theta>t$ تابع $H(\theta)=\theta^{-(N+M)}(\theta-t)^{Z}$ تنها یک ماکزیمم در نقطهٔ $\dot{\theta}=\frac{N+M}{N+M-Z}t$ دارد و به ازای $\theta\geq\dot{\theta}$ بطور یکنوا نزولی است و بنابراین اگر $\dot{\theta}>Y_{\max}$ ، تابع درستنمایی (۷) در $\dot{\theta}$ ماکزیمم و اگر $\dot{\theta}<Y_{\max}$ ، در $Y_{\max}$ بیشینه خواهد شد؛ یعنی:

$$\hat{\theta}=\begin{cases}\dot{\theta} & \dot{\theta}>Y_{\max}\ \&\ Z\geq 1\\ Y_{\max} & \text{در غیر این صورت}\end{cases}$$

امّا چرا حل این مثال به روش الگوریتم EM اشتباه بود؟

دلیل کارا نبودن این روش این است که لگاریتم تابع درستنمایی به ازای تمامی مقادیر $\theta>0$ وجود ندارد، یعنی مقدار مورد انتظار در زیرمجموعه‌ای از فضای پارامتر تعریف نشده است. برای روشن شدن این موضوع فرض کنید در آزمایش دوّم یک لامپ در زمان بقای t مورد آزمایش قرار گرفته و X_m طول عمر مشاهده نشده‌اش باشد، تابع توزیع احتمال غیر شرطی به صورت زیر است:

$$f_X(x_m;\theta)=\begin{cases}\frac{1}{\theta} & 0\leq x_m\leq\theta\\ 0 & \text{در غیر این صورت}\end{cases}$$

در j امین تکرار مرحلهٔ M از الگوریتم باید $\theta^{(j)}$ای که $l^{(j)}(\theta)=E_{\mathbf{X}|\mathbf{Y},\theta^{(j-1)}}[l_c(\theta;\mathbf{Y},\mathbf{X})]$ را ماکزیمم می‌سازد، پیدا شود که این شرطی بودن روی $X_m\mid Y_m$ ، به معنی شرطی بودن روی $X_m>t$ و استفاده از $\theta^{(j-1)}$ به عنوان پارامتر می‌باشد و در نتیجه X_m توزیع یکنواخت روی $[t,\theta^{(j-1)}]$ خواهد داشت.

حال به ازای مقادیر $\theta<\theta^{(j-1)}$ ، ممکن است مقدار $f(x_m;\theta)$ و در نتیجه مقدار تابع درستنمایی صفر شود، و بنابراین $l^{(j)}(\theta)$ ، به ازای $\theta<\theta^{(j-1)}$ وجود نداشته باشد و فرمول (۷) نیز به این موضوع اشاره دارد بطوریکه ممکن است $1-\frac{t}{\max(t,\theta)}$ و در نتیجه $L(\theta)$ صفر شود.

۲. نتیجه‌گیری

این مقاله از تمرینات درس های آمار ریاضی دانشجویان فوق‌لیسانس سرچشمه گرفته است. اولین مثال بر خاصیّت خودسازگاری الگوریتم تاکید دارد و بر خلاف انتظار بسیاری از دانشجویان، در هر موقعیّتی این خاصیّت برقرار نیست.

حتی بهترین دانشجویان هم ممکن است موفّق به حل مثال دوّم نشوند، و در دامی که بوسیلهٔ اوّلین مثال گسترده شده است بیافتند (منظور جایگذاری ساده امید شرطی داده‌های مشاهده نشده بجای داده‌های مشاهده نشده می‌باشد). درسی که از این مثال می‌آموزیم این است که نمی‌توان گفت مرحلهٔ E شامل یک جایگذاری سادهٔ امید شرطی داده‌های مشاهده نشده بجای داده‌های مشاهده نشده است (گر چه این موضوع در مورد بسیاری از کاربردهای مهم الگوریتم EM صادق است)، بلکه در مرحلهٔ E مقدار مورد انتظار لگاریتم تابع درستنمایی کلیهٔ داده‌ها به شرط داده‌های مشاهده شده قرار داده می‌شود. اگر در زیرمجموعه‌ای از فضای پارامتر تابع درستنمایی صفر شود، آنگاه لگاریتم تابع درستنمایی وجود ندارد و در این حالت الگوریتم EM کاربرد ندارد.

مراجع

[1] Dempster, A. P., Laird, N. M., and Rubin, D. B., 1977, Maximum Likelihood from Incomplete Data via the EM Algorithm, (with discussion), *Journal of the Royal Statistical Society,* Ser. B. 39, 1-38.
[2] McLachlan, G., and Krishnan, T., 1997, *The EM-algorithm and Extensions*, New York: Wiley.
[3] Tanner, M. A., 1996*, Tools for Statistical Inference,* (3rd ed.), New York: Springer-Verlag.
[4] Wu, C. F. J., 1983, On the Convergence Properties of the EM Algorithm, *The Annals of Statistics*, 11, 95-103.

شایان ذکر است مقاله حاضر، ترجمهٔ (همراه با توضیح و تلخیص) مقاله زیر می‌باشد:

Flury, B. and A. Zoppe, (2000), Exercises in EM, *The American Statistician*, 54, 3, 207-209.